12. Entrer dans l'anonymat. Etude des "entités dénommantes" dans un corpus oral

Iris ESHKOL (Université d'Orléans - LLL)

0. Introduction

L'enquête Sociolinguistique à Orléans (désormais ESLO1), menée entre 1968 et 1971, a permis la constitution d'un grand corpus de français oral connu sous le nom de Corpus d'Orléans. Les objectifs de ce projet étaient de mettre à disposition un grand corpus de français oral spontané pour des études en linguistique et en didactique. Le Corpus d'Orléans comprend environ 200 interviews référençant les propriétés sociolinguistiques des locuteurs et des situations, soit au total plus de 300 heures de parole (environ 4 500 000 mots) incluant pour moitié des entretiens en face à face et pour moitié une gamme d'enregistrements variés (conversations téléphoniques, réunions publiques, transactions commerciales, repas de famille, entretiens médico-pédagogiques, etc.). En partant des acquis d'ESLO1, une nouvelle enquête, dénommée ESLO 2, a été mise en chantier par le CORAL (Centre Orléanais de Recherche en Anthropologie et Linguistique, EA 3850, devenu Laboratoire Ligérien de Linguistique en 2008). Réunis, ESLO1 et ESLO2 formeront une collection de 700 heures d'enregistrement, soit plus de 10 millions de mots dont l'objectif premier est d'être mis à la disposition de la communauté scientifique et d'un plus large public en rendant ce corpus accessible par internet. La disponibilité effective répond à une préoccupation actuelle forte des recherches en linguistiques.[1]

Dans un premier temps, en 2005, le CORAL a entrepris de reformater le corpus ESLO1 afin de le rendre compatible avec les méthodes et les pratiques actuelles. Cette exploitation du corpus se découpe en sept étapes : numérisation, gestion du corpus et des métadonnées (les données qui permettront l'identification du contenu, la description complète de son mode de production et les caractéristiques techniques du format), transcription synchronisée sur le signal avec une sortie au format XML, étiquetage, anonymisation, interface de requête, diffusion. Diffuser le Corpus d'Orléans selon les techniques actuelles, implique une démarche fondée sur de "bonnes pratiques" juridiques et éthiques. Ce travail a été fait en partenariat avec le

[1] Par exemple les projets ASILA, EPML 50, CATCOD, ANR corpus, CLAPI, PFC, ESLO, *Programme Corpus de la parole*.

programme "Corpus de la parole" de la DGLFLF-MCC et du CNRS. Il reprend les préconisations de l'ouvrage *Corpus oraux, Guide des bonnes pratiques 2006* émanant d'un groupe de travail constitué de linguistes, juristes, informaticiens et conservateurs. Ainsi, si pour des analyses scientifiques précises, le corpus brut reste le seul objet d'analyse possible, la diffusion par internet requiert un corpus anonymisé. L'objectif du projet de construire un portrait sonore de la ville et de ses habitants implique à un haut degré des propos dont la diffusion demande une extrême prudence (informations personnelles, confidences, avis exprimés, etc.). L'état de l'art en 1970 n'incitait pas à s'assurer que les personnes enregistrées étaient disposées à donner leur autorisation pour une exploitation de leurs paroles.

Bien que l'on parle souvent d'anonymisation, la question légale concerne principalement l'assurance qu'il sera impossible d'identifier des personnes. Juridiquement l'anonymisation sert à qualifier l'opération par laquelle se trouve supprimé dans un ensemble de données, recueilli auprès d'un individu ou d'un groupe, tout élément qui permettrait l'identification de ces derniers. Le nom propre n'est donc pas le seul élément qu'il faut prendre en compte. On pourrait parler de "dépersonnalisation" des données, comme le fait la loi fédérale allemande sur la protection des données à caractère personnel du 23 mai 2001. (Olivier *et alii* 2006). Bien sûr il ne s'agit pas de rendre totalement impossible l'identification d'un locuteur (il faudrait alors brouiller la voix sur l'ensemble de l'enregistrement, ce qui rendrait toute analyse linguistique impossible) mais il convient de concevoir des corpus aux formes variables et adaptables à différents contextes d'exploitation. Ce travail qui permet d'échapper à l'alternative entre données disponibles et données inaccessibles nécessite de repérer et de classer les données porteuses d'identification en leur attribuant le cas échéant "une charge identificatrice potentielle" relative aux contextes de production et d'exploitation.

Dans cet article, nous nous interrogerons sur le rôle qu'occupent ces éléments dans le discours oral, le rôle des noms propres dans le processus d'identification, les différents processus d'identification que permet le langage, la diversité lexicale de ces éléments, les différentes informations selon lesquelles on peut identifier une personne et enfin sur la possibilité du traitement automatique de ces éléments dans le cadre de l'anonymisation des corpus oraux.

1. Plusieurs stratégies d'identification du locuteur

Selon le *Dictionnaire d'analyse du discours,* "l'identité résulte, à la fois, des conditions de production qui contraignent le sujet, conditions qui sont inscrites dans la situation de communication et/ou dans le préconstruit

discursif, et des stratégies que celui-ci met en œuvre de façon plus ou moins consciente" (Charaudeau & Maingueneau 2002 : 300). Les auteurs distinguent une identité psychosociale consistant en traits qui définissent le sujet selon son âge, son sexe, son statut, etc. et une identité discursive du sujet énonciateur "qui peut être décrite à l'aide de catégories locutives, de modes de prise de parole, de rôles énonciatifs et de modes d'interventions" (*ib.*) Nous n'allons pas nous intéresser, dans cette étude, aux stratégies discursives que choisit le sujet parlant pour se construire une identité : sa manière de prendre la parole, de thématiser ses propos, d'organiser son argumentation. De la même manière, nous laisserons de côté, pour le moment, d'autres strates de l'identification du locuteur dans le discours :

- l'une où le locuteur ne s'identifie pas du tout (les vérités générales, les descriptions "objectives"...)
- les déictiques qui "concourent à situer l'énoncé" (Dubois 1973 : 137) et qui renvoient "à la situation spatio-temporelle du locuteur ou au locuteur lui-même" (Rey, *Dictionnaire historique de la langue française*)
- des indications sur la personne par des marques lexicales et morphologiques (le féminin pour une femme, le tutoiement de l'enquêteur qui indique un degré de familiarité...).
- la voix

Notre objectif est d'étudier des éléments dans le discours du locuteur permettant son identification par un éventuel utilisateur du corpus. Nous appelons ces éléments *entités dénommantes*. Les entités dénommantes servent à identifier le locuteur en le mentionnant par son nom ou en représentant certains de ses traits ou de son quotidien. Dans la suite de cet article, nous préciserons cette définition, ainsi que la nature des éléments identifiants.

2. La notion d'*entité dénommante*

Il est difficile de rendre compte du mécanisme cognitif en jeu dans le processus de reconnaissance d'un individu. Prenons l'exemple de la reconnaissance des visages où il n'est pas toujours si aisé de décrire verbalement ce qui a conduit à la reconnaissance d'une personne et quels critères y ont contribué. Picoche (1986 :129) dit au sujet des verbes *savoir* et *connaître* :

> "il est parfaitement possible de dire d'une part : Je connais cet enfant-là, je le connais bien, même, mais je ne sais ni son adresse, ni sa date de naissance, ni même son nom de famille, et d'autre part, je sais que cet enfant s'appelle Paul Dupont, qu'il habite rue Victor-Hugo, qu'il est né en 1957, que ses parents sont postiers, […], mais à vrai dire, je ne le connais pas."

Le processus d'anonymisation du corpus va en quelque sorte simplifier la tâche, car on cherchera des indices, des traces "visibles" qui permettront d'identifier le sujet parlant dans le discours. La reconnaissance du locuteur semble passer par la connaissance de certaines de ses propriétés caractéristiques. On peut supposer qu'une entité dénommante (*nom rare, handicap, caractéristique particulière*) ou une série de ces entités (*nom, métier, lieu de travail, loisir, etc.*) est associée à un individu particulier dans la mémoire à l'aide d'un certain lien dénominatif qui sera réactivé lors de leur apparition dans le discours. Il faut prendre en considération les facteurs contextuels qui entourent l'énonciation de ces entités. C'est le contexte qui permettra de réduire le champ d'application de ces éléments à un seul porteur, de le distinguer des autres référents possibles comme dans le cas d'utilisation des noms propres au lieu du prénom ou de l'anthroponyme seuls ou d'une description de l'individu. L'identification peut se faire grâce aux connaissances que l'utilisateur du corpus maîtrise concernant le locuteur ou la personne mentionné dans le discours de celui-ci.

Plusieurs études linguistiques, parlent des traits caractéristiques et définitoires d'un objet. Nous avons cité Charaudeau qui distingue les deux types d'identité du sujet. En sémiotique narrative, Hamon (1977), utilise le terme de " qualification différentielle" pour la série de traits indicateurs de l'importance des personnages dans les romans (descendant d'une famille représentée dans le même roman, porteur d'un nom propre, la description physique et psycho-sociologique, etc.) qui distinguent ce personnage des autres. Ces derniers ne doivent pas se voir attribuer ces propriétés ou seulement partiellement. Ce fait permet à ce personnage d'être considéré comme le héros.

En linguistique cognitive, Jonasson (1994 :141) déclare à propos des noms propres:

> "la description des particuliers désignés par les Npr est donc en partie comparable à celle des classes dénotée par les Nc. Elle comporte l'insertion de l'item dans une taxinomie, suivie de la spécification des

> propriétés typiques permettant son identification et éventuellement accompagnée de l'indication des associations culturelle".

Selon nous, les entités dénommantes sont les éléments descriptifs qui permettent de distinguer le sujet parlant des autres et, par conséquent, de le reconnaître. Pour identifier le locuteur, il suffit de le nommer (s'il s'agit d'un nom rare) et/ou de le décrire par certaines de ses caractéristiques. Les noms propres font donc partie de ces entités au même titre que des noms communs. La distinction entre nom propre et nom commun n'est pas pertinente dans le cadre de notre étude par le fait même que les noms propres et les noms communs fonctionnent de concert pour désigner un référent.

Des études linguistiques ont montré (Jonasson 1994, Leroy 2004) qu'il est difficile de distinguer nettement le nom propre du nom commun quels que soit les critères adoptés (la majuscule, l'impossibilité de traduction, l'absence de l'article, l'incompatibilité avec des déterminants, la mono-référentialité, le manque de sens, etc.).

Sarah Leroy (2004 : 30), critique la thèse de Kripke (1972) qui "s'appuie sur l'idée que les noms propres ne sont pas descriptifs, qu'ils ont une fonction de désignation et d'identification pures, et sont de simples étiquettes posées sur des éléments du réel, ne disant rien de ces éléments. " et met en doute cette notion de "l'unicité référentielle" pour le nom propre en donnant comme exemple :

- des noms propres qui renvoient vers plusieurs référents (*M. Dubois*)
- des noms propres renvoyant à la catégorie des référents

> "*Il y a souvent un Ernest Backes derrière les scoops. Un anonyme blessé, autodidacte, un temps favori des puissants, éjecté sans égards ensuite, qui règle ses comptes au nom d'un combat désintéressé pour la justice et la démocratie*"(p.23)

- des noms propres existant linguistiquement sans désigner des individus réels (*personnage mythologique*)
- des noms communs assurant une désignation unique (*lune*).

> "Ce critère de l'unicité référentielle est donc lui aussi discutable, bien qu'en partie fondé. S'il correspond bien au fonctionnement du nom propre, il ne peut suffire à le définir ni à en délimiter la catégorie, ne serait-ce que parce que le nom commun y répond également dans certains cas, et que le nom propre y échappe dans d'autres cas" (Leroy 2004 : 24)

Nous pouvons ajouter d'autres exemples contredisant la thèse de Kripke :

- les noms propres n'assurent pas seulement l'efficace désignation d'un référent du monde. En désignant quelqu'un par son prénom ou par son statut (*Madame)* on ajoute une information sur ses origines ou son statut civil. Ainsi, les noms propres peuvent aussi décrire un référent du monde.
- l'acte de désignation dépend fortement du contexte de l'énonciation. La prise en compte du contexte s'avère encore plus importante dans le cas du dialogue. Un nom de lieu, par exemple, n'a pas la même valeur s'il est présent dans la réponse sur les origines du locuteur ou s'il se trouve dans la réponse à la question sur les lieux où on parle bien le français. De la même façon, un toponyme faisant partie du nom de l'institution (*Collège de France*) ne renvoie plus vers un lieu.
- enfin, comme l'a mentionné Leroy, les noms communs peuvent également désigner un référent. Il ne s'agit pas seulement, d'une description définie : *le boucher vient tout à l'heure* (si l'on est dans un petit village qui n'a qu'un boucher), mais aussi d'une série de descripteurs qui dans un contexte donné peuvent ensemble, partiellement ou en combinaison avec des noms propres, permettre l'identification du locuteur.

Jonasson (1994 : 138) montre que le nom propre ne peut pas être considéré seulement dans son emploi référentiel. Elle distingue les noms propres "connus", "historiques", des noms propres "familiers" comme *Paul, Marie, etc.* qui "sont souvent associés à de nombreux particuliers […] mais désignent dans un champ restreint (famille, classe, bureau, études, vie privée, village, etc.) un seul ou un nombre limité de particuliers." Jonasson et Kleiber (1981) mentionnent la difficulté d'interpréter ces noms sans ajout d'autres renseignements sur la relation personnelle, par exemple, entre le locuteur et le référent visé (*Paul, c'est mon fils*). On a donc besoin d'une certaine extension du contexte pour pouvoir leur associer un référent.

Ainsi, le repérage du référent se fait à partir de divers types de connaissances (linguistiques, métalinguistiques et encyclopédiques) qui, d'une manière ou d'une autre, relèvent toutes du contexte, puisqu'elles sont supposées être présentes chez les sujets parlants au moment de l'énonciation. Avant de passer à l'analyse du corpus, nous aimerions faire quelques remarques sur le choix du terme « entités dénommantes ».

Selon les dictionnaires, désigner c'est "indiquer de manière à faire distinguer de tous les autres par un geste, une marque, un signe" (*Le Nouveau Petit Robert*, 1993). Dénommer c'est "attribuer un nom à quelqu'un ou à quelque chose" (*TLF* en ligne).

Ainsi, les entités dénommantes désignent le locuteur en le dénommant ou en le décrivant.

Pour Kleiber (1984 : 80), l'acte de dénomination "consiste en l'institution entre un objet et un signe X d'une association référentielle durable" codée, apprise, mémorisée préalablement. Il peut s'agir des noms propres comme des noms communs. La désignation, à son tour, est constituée par le processus qui crée une association occasionnelle entre un signe linguistique X et un élément de la réalité. Elle n'est donc ni codée, ni mémorisée.

En donnant le nom d'entités dénommantes aux éléments permettant d'identifier le sujet parlant, nous n'avons pas repris la définition de Kleiber, même si l'on pourrait dire que les entités dénommantes renvoient vers l'objet du monde réel grâce à l'"association référentielle" durable codée, apprise, mémorisée qu'elles nouent avec. Le nom d'entités dénommantes est utilisé pour le différencier du terme d'*entité nommée* en traitement automatique du langage, où il désigne les noms propres mais aussi les expressions de temps et de quantité.

3. Etude du corpus

Dans cette partie, nous présenterons un test mené sur un sous-corpus d'ESLO1. L'objectif visé est l'étude des entités dénommantes dans le corpus oral afin d'identifier :

- le type et la nature des éléments permettant l'identification,
- la proportion des noms communs et des noms propres parmi ces éléments,
- les problèmes que pose leur repérage automatique ;

Pour étudier des entités dénommantes, nous avons choisi un "corpus-échantillon" composé des transcriptions de 20 entretiens (d'une durée variable d'une à deux heures, soit 11'000 mots en moyenne). Le questionnaire de l'entretien contient tout d'abord des questions préliminaires (*Depuis combien de temps habitez-vous Orléans ? Qu'est-ce qui vous a amené à vivre à Orléans ? Est-ce que vous vous plaisez à Orléans ?* etc.), puis des questions sur le travail et les loisirs du locuteur et des membres de sa famille. Enfin, sont abordés :

- l'enseignement (*Qu'est-ce qu'on devrait apprendre surtout aux enfants à l'école ? Dans quelles matières aimeriez-vous que vos enfants soient forts ?* etc.) ;

- la politique (*Est-ce que, d'après vous, on fait assez pour les habitants d'Orléans ? Que pensez-vous des événements de mai 68 ?* etc.) ;
- la langue et les habitudes culturelles (*Un étranger veut venir en France pour apprendre le français, dans quelle région est-ce qu'il doit aller d'après vous, dans quelle ville ? Quelqu'un frappe à la porte de cette pièce, qu'est-ce que vous lui dites ?* etc.).

3.1. Entités nommées versus entités dénommantes

Dans le cadre du projet VARILING,[2] nous collaborons avec le Laboratoire Informatique (LI) de Tours (Denis Maurel, Marie-Aimée Gazeau). Dans ce cadre, nous utilisons le système de reconnaissance des entités nommés CasSys développé dans le LI par Nathalie Friburger dans le cadre de sa thèse. Ce système permet de réaliser des cascades de transducteurs[3] en utilisant les outils fournis par Unitex. L'un des objectifs de ce test était de vérifier dans quelle mesure l'outil de repérage des entités nommées, avant son adaptation à l'oral et à ce corpus en particulier (actuellement, ce travail est en train de se réaliser avec succès au sein du laboratoire), pouvait être utilisé dans la reconnaissance des entités dénommantes, et quels sont les éléments, attestés ou non, dans les résultats.

Le corpus de travail a donc été intégralement traité au moyen de ce logiciel. La deuxième phase du travail a consisté à vérifier les résultats. Les étudiants du master ILTC de l'université d'Orléans ont effectué cette partie de la tâche. Ils ont reçu les fichiers résultats contenant l'annotation des entités nommées relevées.

> *{S}RC: depuis combien de temps habitez vous <lieu val=\"top\"><nom>Orléans</nom></lieu> ?*
> *{S}GJ 131: oh ça fait neuf ans depuis dix neuf cent soixante*
> *{S}RC: vous vous plaisez à<lieu val="pays"><nom> Orléans</nom></lieu> ?*
> *{S}GJ 131: oui et non*
> *{S}RC: [rire] pourquoi ça ?*
> *{S}GJ 131: bah parce que j'ai j'ai toujours euh je suis <lieu val=\"top\"><nom>Lorrain</nom></lieu> alors j'ai je suis né*

---

2 VARILING Projet ANR 2006.

3 Un transducteur est un automate à nombre fini d'états dont les transitions sont étiquetées par un couple de symboles : un symbole reconnu en entrée et un symbole produit en sortie. "Une cascade de transducteurs est une succession de transducteurs appliqués sur un texte, dans un ordre précis, pour le transformer ou en extraire des motifs." (Friburger 2002 : 49). Chaque transducteur utilise les résultats des transducteurs précédents.

en<lieu val="en"><nom> Lorraine</nom></lieu> et puis j'ai toujours été en<lieu val="en"><nom> Lorraine</nom></lieu> et je préfère la<lieu val="pays"><nom> Lorraine</nom></lieu> à l'Orléanais[4]

Il a ensuite été demandé d'étudier l'annotation effectuée pour marquer en premier lieu deux types de balisage :

- balisage correct qui correspond aux entités nommées portant des informations jugées pertinentes pour l'étude ;
- balisage inutile ou erroné qui ne correspond pas aux objectifs.

Des exemples de balisage inutile concernent par exemple les noms de lieux comme *Orléans*, *France*, des noms de présentateurs de télévision ou encore les noms de fêtes comme *Pâques*, etc. qui n'apportent pas d'informations particulières sur l'identité du locuteur.

Les étudiants devaient, en second lieu, trouver et annoter les éléments absents dans le balisage effectué automatiquement et qui semblaient révéler l'identité du locuteur et/ou de ses proches. La figure 1 montre la répartition entre les trois types de balisage :

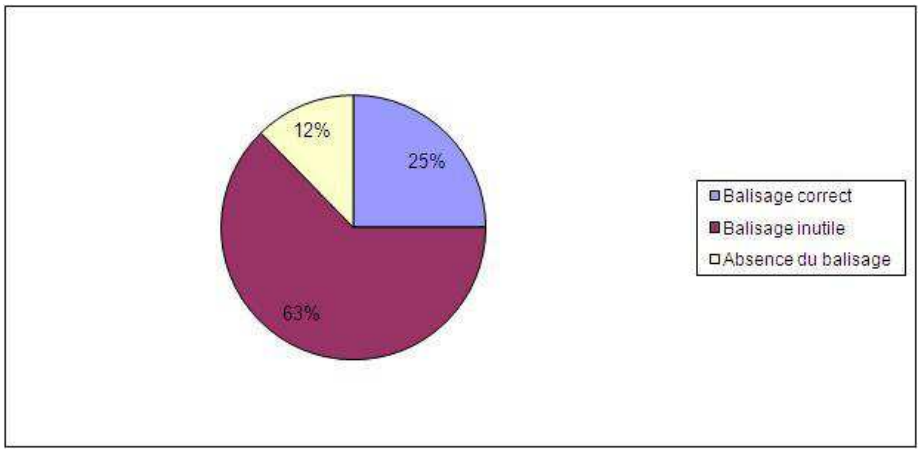

Figure 1

[4] L'annotation fait par cet outil est basée sur la typologie des noms propres effectuée dans le cadre du Projet Prolex : http://www.cnrtl.fr/lexiques/prolex/. Pour la description de cette typologie voir (Tran & Maurel 2006).

Les entités dénommantes résultent de la somme du balisage correct et de l'absence de balisage. Ce pourcentage montre clairement que le traitement automatique des entités dénommantes ne peut pas se limiter à la reconnaissance des entités nommées car, d'une part, ces derniers ne jouent pas toujours le rôle d'éléments identifiants et, d'autres part, ces éléments ne peuvent pas couvrir la totalité des entités dénommantes dont une partie relève des noms communs et n'ont aucun rapport avec des entités nommées. La figure 2 présente justement la proportion des noms propres et des noms communs parmi les entités dénommantes :

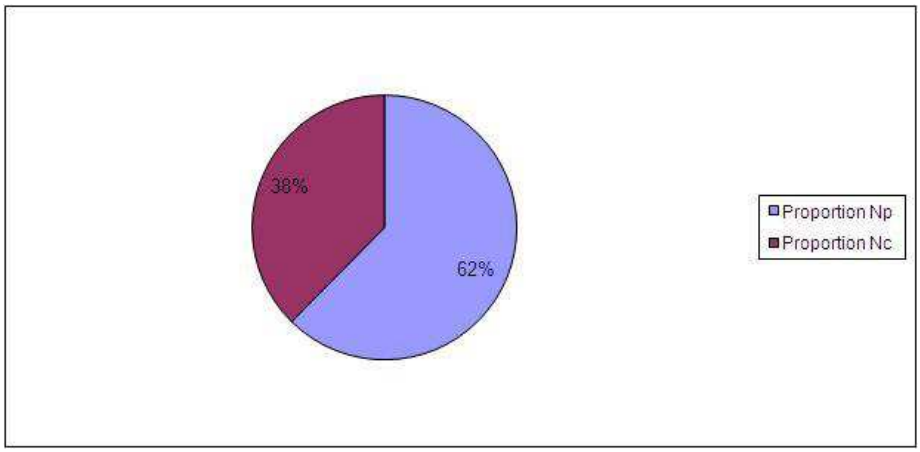

Figure 2

Ainsi, les noms communs occupent le tiers des entités dénommantes et doivent être pris en compte en même titre que les noms propres. Cependant, la proportion des noms propres reste très importante parmi les entités dénommantes, c'est pourquoi l'utilisation d'un outil de reconnaissance d'entités nommées comme CasSys est très utile dans notre tâche. Il présente l'avantage d'être facilement adaptable au domaine traité car le linguiste peut lui-même créer et modifier les graphes selon le corpus auquel il est confronté. La partie des noms communs absents pourrait être reconnue grâce aux graphes spécialement créés. Pour le reste, il n'y a pas d'autre solution qu'un travail manuel. Par contre, il reste un problème à résoudre en ce qui concerne le nombre très élevé d'entités nommés ne jouant aucun rôle dans

l'identification du locuteur mais donnant des informations générales et neutres. Le rôle du contexte semble ici la meilleure piste de solution[5].

3.2. Typologie des entités dénommantes

Ce test nous a permis également d'établir l'esquisse d'une typologie des entités denommantes selon les informations qu'elles fournissent et leur structure interne.

Le corpus d'entretiens "face à face" est un corpus riche d'informations personnelles car il est composé des entretiens où la partie du questionnaire porte sur l'identité du locuteur : son travail, ses études, ses loisirs, sa famille. Ce choix nous a permis de travailler plus facilement sur le processus d'identification grâce à la richesse de ce type d'information dans le corpus. Ainsi, lorsque le locuteur parle de la profession qu'il exerce, il emploie différentes formes :

- le nom direct du métier ou son synonyme : *professeur des écoles ou institutrice,*
- le contenu, la description de son travail : *j'enseigne les maths,*
- le lieu : *je travaille au collège de Saint-Jean-de-Braye.*

Il est très difficile de créer une typologie de ces éléments qui sont de nature hétérogène. A cela s'ajoute la qualité subjective de cette opération. Dans le travail que nous avons effectué avec les étudiants, nous avons pu nous rendre compte de la variété d'interprétation selon les annotateurs, la même information pouvant ne pas être traitée de la même manière selon les étudiants et des variations pouvant être présentes dans les choix successifs qu'effectue un même annotateur.

Dans ce qui suit, nous allons essayer de décrire ces éléments dans le but d'obtenir un classement envisageable sans pour autant construire des groupes homogènes.

3.2.1. Types d'identifiants

Corblin (1983 : 204) distingue deux types d'anaphores : les désignateurs essentiels qui renvoient à des propriétés intrinsèques de l'entité référentielle et les désignateurs contingents et éphémères qui opèrent une "saisie ponctuelle de caractéristiques purement conjoncturelles du référent". On utilisera la même démarche pour parler du processus d'identification.

---

[5] Il en sera parlé dans la partie consacrée au rôle du contexte dans la tâche de l'anonymisation.

Le processus d'identification peut être direct ou indirect d'où la distinction entre :

- identifiant direct (unicité référentielle): il permet, à lui seul, de distinguer un individu des autres et renvoie directement vers un référent unique; sa présence est nécessaire et suffisante pour la reconnaissance de l'individu. Le processus n'est pas progressif, il est ponctuel :
    - nom rare de la personne, surnom
        - *Ostrowetsky*
        - *dans ma classe quelquefois ils ne sont pas obéissants …on m'appelle la maîtresse des fous*
    - métier rare (*général*) ou statut (*maire*)
    - caractéristique rare (*nombre élevé d'enfants, handicap*)
        - *mon père a fondé le plus grand cabinet d'ophtalmologiste de la ville*

- identifiant non direct : sa présence seule ne permet pas l'identification, mais en combinaison avec d'autres identifiants, il peut désigner un référent unique. Par exemple :

> *le locuteur est patron d'un bar au moment d'enregistrement, et avant il travaillait dans l'aviation militaire*

Il s'agit d'attributs qui ne sont pas uniques et peuvent être partagés par plusieurs individus. Le processus d'identification est progressif, il se construit au fur et au mesure de l'accrétion des indices. Parmi ces identifiants non directs, on distinguera ceux qui sont les plus sensibles à l'anonymisation et ceux qui apportent une information plus importante et plus spécifique, de ceux qui sont plus généraux. Cette distinction nous permettra d'opposer

- les noms de famille comme *M.Dupond* ou *M.Durand* aux autres
- les noms de métiers *professeur de physique / enseignant*

3.2.2. Diversité lexicale et sémantique des entités dénommantes

Les entités dénommantes peuvent être de nature lexicale et sémantique très variée. On ne peut donc pas s'arrêter à la seule catégorie nominale.

Tout d'abord, nous citerons des entités nommées "classiques" repérables automatiquement :

- noms de personne :
    - *patronnée par Suzanne Fouché*

        - *et puis alors vous avez aussi euh le recteur Antoine*
- noms de lieu :
        - *dans l'Indre*
        - *je suis allée en Espagne*
- noms d'organisation :
        - *euh je suis je travaille à l'hôpital d'Orléans quoi*
        - *je fais parti de la SPA*
        - *le collège de Saint-Jean de la Ruelle*
        - *usine Michelin, collège Benjamin Franklin, résidence Dauphine*
        - *école normale Bellegarde*

Ces entités, si elles ne sont pas repérées dans le dictionnaire des noms propres, comportent souvent un nom commun descriptif (*hôpital*, *collège*, *ville etc. de*), un mot déclencheur, dont le sens indique à quelle catégorie notionnelle elles appartiennent, ce qui facilite la construction des patrons et, par conséquent, leur reconnaissance et annotation automatique.

- éléments chiffrés : âge, année
        - *en mille neuf cent soixante-neuf Mademoiselle*
        - *j'avais une fille de quinze ans*
        - *j'ai mon fils qui a vingt-huit ans*

Malheureusement, la reconnaissance des entités nommées ne suffit pas à repérer toutes les informations concernant le sujet parlant car le discours du locuteur contient souvent d'autres éléments permettant de l'identifier par recoupement :

noms de métiers

je suis enseignant dans dans l'école publique
comme officier j'ai été obligé de rester 45 ans

origine

oui je suis orléanaise

maladies

j'ai une maladie du foie
ça lui a même occasionné une petite scoliose déformation légère de la colonne vertébrale

Le cadre de l'entretien influence l'apparition de certaines informations qui dépendent majoritairement des questions posées. Ainsi, les entités

dénommantes du corpus peuvent être classées dans les rubriques suivantes : *origine, travail, études, famille, âge, loisirs* :

- origine
  oui je suis orléanaise
  et je suis née ici dans la maison
  d'origine je suis dauphinien

- travail
  actuellement j'enseigne à côté de Châteauroux et j'étudie à Orléans
  mon travail d'élève infirmière ou le travail d'auxiliaire de puériculture
  je m'occupe uniquement de malades adultes puisque je fais mes études d'infirmière
  je suis à la Source je travaille au laboratoire
  j'enseigne les mathématiques
  j' ai un frère qui est à l' armée à Tours en ce moment
  mon fils est venu manger qui est soldat
  mon père qui étant psychologue

- études
  je suis licencié licencié en physique
  j'étais interne un an et demi pensionnaire tout le reste de mon existence de lycéen
  j'ai fait 4 ans de faculté

- loisirs
  je suis scout de France
  je suis animatrice louvetisme sur Orléans
  je fais beaucoup de photographies
  je milite dans un mouvement de jeunesse
  le jeudi soir où j'anime un un atelier photos
  je fais partie de chorale oui je fais euh à la chorale à la cathédrale

- famille
  mon neveu […] Jean-Pierre
  j' ai perdu le père très jeune celle-là elle avait douze ans
  et puis malheureusement j'ai perdu mon mari assez tôt aussi et ça fait quinze ans

Ces thèmes peuvent se chevaucher comme on voit dans les exemples :

j' ai un frère qui est à l' armée à Tours en ce moment
mon fils est venu manger qui est soldat
mon père qui étant psychologue

où on trouve les informations sur le travail et/ou l'âge des membres de la famille.

Chaque type d'information peut être présenté à travers un groupe nominal ainsi qu'avec des expressions plus étendues. Ce passage se manifeste par l'ajout de propriétés supplémentaires à la classe présentée par le groupe nominal minimal, ce qui diminue l'extension de la classe et rapproche le groupe d'une référence plus individualisante. Prenons comme exemple, le domaine du travail, lorsque le locuteur essaie de le décrire et de donner plus de détails sur ses fonctions :

Je suis enseignant dans l'école publique =>
- Je suis maître auxiliaire, j'enseigne des mathématiques modernes des mathématiques classiques de la chimie et de la technologie
- actuellement j'enseigne à côté de Châteauroux et j'étudie à Orléans

En tant que mécanicien =>
- je travaille dans l'usine qui est juste à côté de la maison là on fait des appareils ménagers

Je suis professeur d'éducation physique=>
- j'enseigne l'éducation physique dans toutes les classes de la sixième à la troisième
- je dispose d'un d'un plateau d'éducation physique qui comporte deux terrains de basket un terrain de hand-ball
- je suis au collège de Saint-Jean-de-Braye

mon travail d'élève infirmière ou le travail d'auxiliaire de puériculture=>
je m'occupe uniquement de malades adultes puisque je fais mes études d'infirmière

Le locuteur nomme d'abord son métier (à cause des questions posées) et ensuite le spécifie. Suite à ces exemples précis, on constate que la spécification du travail s'effectue souvent par :

- les verbes d'activité : *s'occuper de, faire de, enseigner* etc.+domaine d'activité

- la précision du lieu de travail : entité nommée introduite par *être* avec fonction locative ou par une préposition locative

Il reste, enfin, des informations difficiles à cataloguer :

- *on a monté une association d'élèves infirmières*
- *nous louons une villa à Royan*
- *mon père a fondé un le plus grand cabinet d'ophtalmologiste de la ville*
- *j'attends un deuxième bébé*

Ainsi, on peut observer une certaine régularité de structures syntaxiques (constructions attributives, les verbes d'occupation, noms de relations familiale, etc.) qui rend possible la création de patrons pour une reconnaissance automatique. Cependant, tout n'est pas repérable automatiquement. La catégorie des informations occasionnelles semble "imprévisible" en raison de son manque d'homogénéité et elle ne peut être repérée que par l'analyse manuelle du corpus, ce qui représente un travail considérable.

La multitude d'éléments personnels, biographiques dans notre corpus soulève une autre question concernant leur pertinence. Sont-ils tous identifiants ? Lesquels devons-nous retenir comme pertinents pour notre tâche ? Est-ce que l'année d'arrivée à Orléans, l'âge des enfants, le lieu de naissance (autre que le Loiret), la nationalité (si elle est française), etc. sont susceptibles de révéler l'identité du locuteur ou de ses proches ? Actuellement, nous n'avons pas de réponses à ces questions. L'annotation de ces éléments paraît dépendre du contexte pris au sens le plus large.

Quand le métier du locuteur est mentionné plusieurs fois dans le discours de façon différente, faut-il masquer à chaque fois l'information ? Prenons l'exemple d'un enseignant :

> *je prépare mes cours , pendant la classe je n'en fais pas , les élèves nous disent , trois ou quatre fois dans le d- dans les copies je trouve la même faute d'orthographe je finis par me dire , moi j'ai de élèves qui ont trente-six heures de cours , c'est difficile pour mes élèves , dans mon travail et bien c' je crois la réussite des élèves , quand je dis aux élèves je suis très contente vivement que nous rentrions en classe , ben voyez là nous avons fait des changements là demandés par les un peu par les élèves aussi à ce qu'il n'[y] ait plus de compositions à ce qu'il n'[y] ait plus de notes*

Les limites de la tâche ne sont pas faciles à fixer, si l'on s'assigne pour objectif de conserver le plus d'informations possible. On pourrait créer une ontologie avec les termes associés à chaque domaine. Ainsi, pour l'enseignement les mots comme *cours, élève, copie, classe, etc.* pourraient être pris en compte et donc annotés automatiquement. Mais si l'on doit effacer toutes les marques de professions, on modifie tellement le corpus qu'il devient peu compréhensible à proportion de la perte d'informations significatives.

La figure 3 montre le pourcentage des différents types d'entités dénommantes présentes dans notre corpus. Il s'agit de noms de personne (*Suzanne Fouché*), de lieux (*Tunisie*), d'organisation (*je fais partie de la SPA ; et si vous n'étiez pas euh dans l'administration euh militaire*), d'activités (*c'est-à-dire que je m'occupais de ce point de vue des familles de militaires ; euh je fais partie du conseil scolaire*), de métiers (*eh bien parce que mon mari étant officier de carrière*) ainsi que d'autres types d'informations personnelles (*j'ai beaucoup déménagé j'ai déménagé quatorze fois ; la vie que vous avez menée euh aux colonies*). Précisons que la distinction entre les catégories : métier, activité et autres n'est pas toujours évidente.

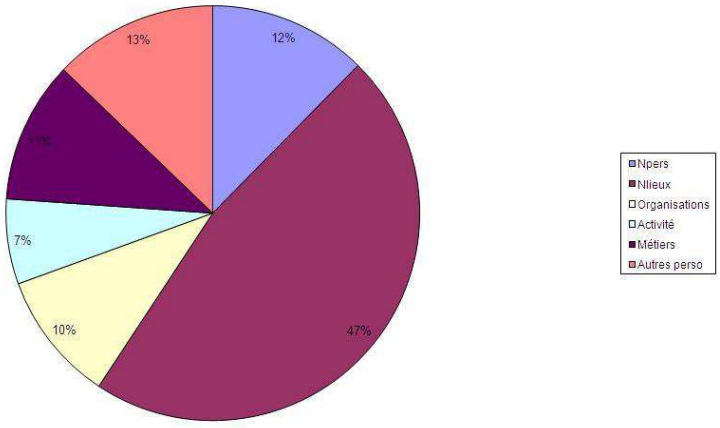

Figure 3

Les toponymes semblent les plus fréquents. Ils représentent la moitié des entités dénommantes. Si l'on compare ces résultats avec le repérage des entités nommées par CasSys, on voit que les noms de lieux sont majoritaires 57% (Fig 4).

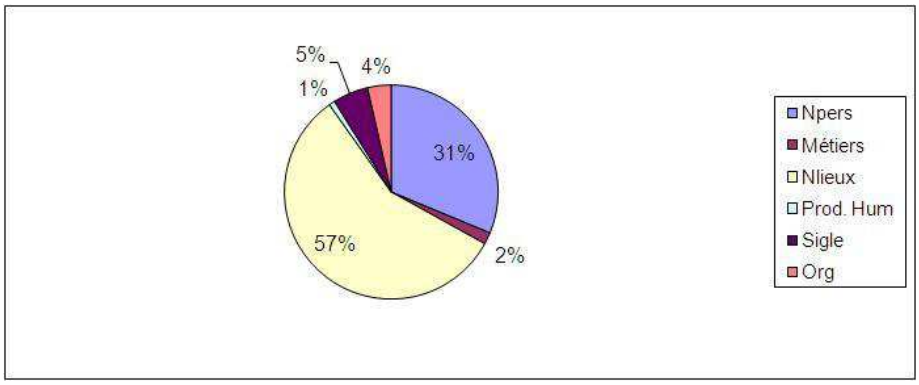

Figure 4[6]

Ce phénomène semble représentatif des corpus oraux.[7] La question qui reste à résoudre, concerne leur rôle parmi les éléments identifiants. Est-ce que les noms de lieux apportent plus d'informations personnelles que les autres éléments ? La précision du lieu de travail, de l'activité, d'une organisation, etc. diminue le nombre de référents possibles et se rapproche d'une référence plus individuelle. Les noms de lieux jouent dans ce cas le même rôle que des modifieurs.

Il est intéressant de noter que les noms de personne ne sont pas très fréquents parmi les entités dénommantes (12%) contrairement aux entités nommées. Ils occupent la deuxième place après les toponymes (31%), se répartissant d'une manière plus au moins équilibrée par rapport aux autres informations (métier, organisation, etc.) (Fig 3). Cette observation montre que la présence du nom de la personne ne suffit pas pour la reconnaître et ne donne pas toujours des informations permettant de l'identifier.

---

6 Nous n'avons pas gardé les mêmes distinctions. Nous avons catégorisé les produits humains comme non représentatifs des entités dénommantes et les sigles comme des organisations.

7 Selon la communication orale de Anne Dister au colloque Nomina 2007.

3.3. Le rôle du contexte

Corblin et Gardent (2005:15) notent que

> "Pour l'analyse linguistique, le contexte pertinent est celui qui recouvre l'ensemble des éléments impliqués par l'activité langagière : les connaissances lexicales et encyclopédiques des participants, la situation physique d'énonciation (participants, lieu, temps) et le contexte linguistique, c'est-à-dire une trace du texte ou du dialogue précédant l'énoncé considéré. Chacun de ces éléments peut à la fois participer à la détermination du sens en contexte et contraindre la forme de l'énoncé produit."

Siblot (2007 : 34) précise que l'acte de parole de la nomination, comme tout acte de parole, doit être envisagé dans son contexte de production et de communication et appréhendé dans le procès d'actualisation. Il importe de tenir compte de cette "contingence historique, culturelle, sociale et individuelle" dont dépendent l'acte de parole et la production du sens en général.

L'acte d'identification dépend fortement du contexte de l'énonciation. En premier lieu, on peut mentionner le contexte immédiat (gauche et/ou droite) d'une entité. Ainsi, le nom de lieu n'aura pas de grand intérêt employé seul, mais employé avec des verbes comme *venir de*, *travailler à* ou avec des noms comme *collège*, *hôpital*, *etc.* il devient identifiant du lieu de travail, d'études ou d'origine de la personne. Ce contexte est largement utilisé pour la reconnaissance automatique des éléments par une approche linguistique fondée sur la description syntaxique et lexicale des syntagmes recherchés. On crée des règles de grammaire ou des patrons décrivant le syntagme et son contexte immédiat en utilisant des marqueurs lexicaux (mots déclencheurs), des dictionnaires de noms propres et des dictionnaires spécifiques (par exemple dictionnaire des métiers). Ces indices permettent de repérer un élément mais aussi de le catégoriser. Ce contexte peut être aussi défini par la question posée. On sort ce faisant des limites de l'énoncé pour étudier un contexte plus large. Le nom de lieu, par exemple, n'est pas signifiant s'il est utilisé pour répondre à la question : "où parle-t-on le mieux le français ? ", par contre il devient un identifiant dans les réponses aux questions concernant les origines du locuteur, ou dans les énoncés décrivant l'emploi du locuteur, pour autant que celui-ci indique le lieu de son travail. De la même manière, les réponses aux questions sur les émissions de télé, par exemple, n'apportent pas d'information personnelle et les noms de personnes qui apparaissent n'ont pas à être pris en compte :

monsieur Fouchet Christian Fouchet ministre de l'Education Nationale
il a surpris beaucoup de personnes Edgar Faure certainement
j'ai entendu parler euh Michel Couaste on dit ç- ça ? on dit on prononce comme ça Michel Couaste?

On pourrait, en partant de là, opérer une distinction entre les questions sensibles dont les réponses peuvent contenir certaines informations personnelles :

Qu'est-ce que vous faites comme travail?
- en quoi est-ce que ça consiste/c'est quoi au juste?
Et votre femme, est-ce qu'elle travaille aussi? Pourquoi (pas)?
Et vos enfants, que font-ils?/ métier?
Qu'est-ce que vous faites de votre temps libre - soirées, week-end?
Comment a-t-on choisi dans votre cas personnel entre l'école publique et l'école libre?
Etc.

et les questions neutres où la présence des entités nommées ne renvoie pas nécessairement au locuteur :

A votre avis, qu'est-ce qu'on devrait apprendre surtout aux enfants ? à l'école ?
Qu'est-ce que vous pensez du latin ? à l'école ?
Pour revenir à la ville d'Orléans, est-ce que, d'après vous, on fait assez pour les habitants d'Orléans?
Ecoutez-vous la radio ? nombre d'heures par semaine/jour ? Votre chaîne préférée ? Vos émissions préférées ?
Etc.

Seules les questions sensibles requièrent d'être prises en compte si l'on veut distinguer l'information neutre de l'information personnelle.

L'entité dénommante repérée doit être étiquetée selon le contexte. Dans la phrase *je travaille au collège de Saint-Jean-de-Bray,* l'entité *collège de Saint-Jean-de-Braye* ne réfère plus seulement à un établissement scolaire en général, c'est une référence à un lieu de travail. Les questions posées pourront donc jouer un rôle important dans la catégorisation adéquate d'une entité repérée.

Enfin, il est nécessaire de prendre en compte le contexte socioculturel de l'époque. Ainsi, les destinations de vacances peuvent être prises en compte car en 1968 très peu de gens voyageaient à l'étranger :

> *j'ai vu aussi pas mal de pays j'ai vu l'Espagne le Portugal euh l'Allemagne l'Italie la Sicile qui m'a beaucoup plu également le la Yougoslavie*
> *nous sommes allés par bateau jusqu'au Cap Nord et retour euh par euh jusqu'à la frontière finlandaise jusqu'à Oslo après nous avons vu euh la Suède et le Danemark Canaries et retour par Dakar*

Certaines informations doivent être parfois déduites du contexte comme dans l'exemple suivant:

> *BV: y a longtemps que vous êtes à Orléans ?*
> *MS530: euh oui euh vingt-deux ans*
> *BV: ça fait euh vous êtes née à Orléans*
> *MS530: oui*

La prise en compte du contexte socioculturel de l'époque et la déduction de nouvelles informations montrent les limites du traitement automatique sur des corpus qui ne correspondent pas à la situation présente. La difficulté réside dans la description formelle de ce type de connaissances qui fait partie des recherches en intelligence artificielle, mais qui s'exerce moins dans le domaine des corpus oraux.

La figure 5 montre des entités dénommantes d'un enregistrement annotées manuellement et extraites grâce aux feuilles de style XSLT sous forme de tableau :

Numéro de l'interview	Travail					Famille		Origine	Etudes
	Lieu	Duree	Institution	Activites	Metier	Lien_famille	Travail		
98									
	près de Montargis à Bellegarde	un an	un collège d'enseignement général	je n'enseigne que l'éducation physique	prof d'éducation physique	femme	préparatrice en pharmacie	je suis arrivé à Orléans en cinquante huit	j'ai fait auparavant un an au au collège Benjamin Franklin
	Bellegarde		collège de Saint-Jean-de-Braye				travaille à la pharmacie		j'ai fait mes années à l'école normale
	Saint-Jean-de-Braye								

Figure 5

Comme l'annotation a été réalisée manuellement, tous les types de contexte ont été pris en compte. Nous avons distingué 4 grands domaines : *travail, famille, origine, études,*[8]chacun pouvant être précisé (lieu, durée, etc.). De cette manière a été créée une fiche individuelle du locuteur. A l'analyse de cette fiche, les informations plus ou moins sensibles à l'anonymisation apparaissent. Ainsi, nous prenons en compte, au nombre des informations, qu'il s'agit d'"un enseignant d'éducation physique au collège de Saint-Jean-de-Braye dont la femme est préparatrice en pharmacie". De telles indications sont probablement suffisantes pour identifier le locuteur. Inversement, son origine, ses études et même le nom de son métier - "enseignant", sans préciser "d'éducation physique"- ne permettent pas de déterminer qui il est. On en conclut que le processus d'anonymisation ne peut pas s'effectuer en une seule étape. Il s'agit d'un traitement décomposable en plusieurs paliers, le dernier ne circonscrivant que les informations les plus « identifiantes », celles qui devront être masquées. Dans ce cas, la suppression de l'indication de lieu devrait suffire, à elle seule, à satisfaire l'anonymisation.

4. Conclusion

L'anonymisation est une étape souvent nécessaire et toujours délicate d'un traitement du corpus oral. On a présenté, dans cet article, un test effectué sur un corpus "sacrifié" afin de pouvoir définir et décrire les éléments permettant l'identification du locuteur. Nous avons appelé ces éléments les *entités dénommantes* par opposition aux *entités nommées* qui ne répondent pas aux mêmes critères. Ainsi, le traitement automatique des entités dénommantes ne peut pas se satisfaire des techniques linguistiques éprouvées de repérage des entités nommées car ni le critère de majuscule, ni les règles utilisant des mots déclencheurs, des dictionnaires de noms propres et des dictionnaires spécifiques ne sont suffisants eu égard à la complexité de la tâche. D'une part, les entités dénommantes dépassent par leur diversité les entités nommées et, d'autre part, les entités nommées repérées doivent fournir des informations sur le locuteur, ce qui n'est pas toujours le cas. Actuellement, Marie-Aimée Gazeau et Denis Maurel (Laboratoire Informatique de Tours) mènent à bien le travail de reconnaissances des éléments identifiants dans le corpus ESLO1. Ils utilisent le logiciel CasSys et adaptent les graphes de reconnaissances d'entités nommées au corpus en essayant de prendre en compte le contexte des questions posées pour augmenter la pertinence des

[8] Cette liste n'est pas exhaustive car on y pourrait ajouter d'autres titres : les loisirs, vacances, etc.

entités nommées repérés. Ils créent également de nouveaux graphes pour reconnaître d'autres éléments ne faisant pas partie des entités nommées (métiers, origine, etc.). Il reste à valider ces éléments car tout ne saurait être masqué. C'est à la définition attendue des paramètres supplémentaires de filtrage des éléments annotés que nous consacrons une part de notre recherche.

Bibliographie

Baude, Olivier *et alii* (2006) : *Corpus oraux : guide des bonnes pratiques 2006.* Paris et Orléans, CNRS-Editions et PUO.

Baude, Olivier et Eshkol, Iris (2006) : Constitution et exploitation d'un grand corpus de "données situées". Problèmes et solutions pour les Enquêtes Socio-Linguistiques à Orléans (1968-2008), *Troisièmes rencontres fribourgeoises sur la linguistique de corpus appliquée aux langues romanes*, Fribourg.

Charaudeau, Patrick et Maingueneau, Dominique (2002) : *Dictionnaire d'analyse du discours*. Paris, Éditions du Seuil.

Corblin, Francis (1983) : "Les désignateurs dans les romans", *Poétique 54,* 199-211.

Corblin, Francis et Gardent, Claire (2005) : « Contexte et interprétation », *Interpréter en contexte.* Lavoisier, Paris, 15-28.

Dubois, J (1973) : *Dictionnaire de linguistique,* Paris, Larousse.

Friburger, Nathalie (2002) : *Reconnaissance automatique des noms propres. Application à la classification automatique de textes journalistiques.* Thèse de Doctorat, Université de Tours.

Hamon, Philippe (1977) : "Pour un statut sémiologique du personnage", dans *Poétique du récit,* Barthes R. *et alii,* Points-Seuil, Paris.

Jonasson, Kerstin (1994) : *Le Nom propre. Constructions et interprétations*. Duculot, Belgique, Louvain-la-Neuve.

Kleiber, Georges (1984) : "Dénomination et relations dénominatives", *Langages* 76, 77-94.

Kleiber, Georges (1981) : *Problèmes de référence : descriptions définies et noms propres*. Paris, Klincksieck.

Kripke, Saul Aaron (1972) : "Naming and Necessity", dans Davidson, D.&G.Harman (éds,), *Semantics of Natural Language,* Dordrecht.

Leroy, Sarah. (2004) : *Le nom propre en français.* Ophrys, Paris.

Picoche, Jacqueline. (1986) : Structures sémantiques du lexique français. Paris.

Schnedecker, Catherine. (1997) : *Nom propre et chaines de références.* Recherches Linguistiques, n. 21, Paris.

Siblot, Paul. (2007) : "Nomination et point de vue : la composante déictique des catégorisations lexicales", *L'acte de nommer. Une dynamique entre langue et discours*. Paris, Presses Sorbonne Nouvelle, 25-38.

Tran, Mickaël et Maurel, Denis. (2006) : "Prolexbase : Un dictionnaire relationnel multilingue de noms propres", *Traitement automatique des langues*, 47 (3), 115-139.